

# Part-Aware and Robust Deep Learning Models for Vision Applications

**Jiaxu Miao**

**Supervisor:** Prof. Yi Yang

Faculty of Engineering and Information Technology  
University of Technology Sydney

Thesis submitted in fulfilment of the requirements for the degree of  
Doctor of Philosophy

June 2021

# Declaration

I, Jiaxu Miao declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: 

Jiaxu Miao  
June 2021

## Acknowledgements

First, I would like to thank my supervisor, Professor Yi Yang. I could not start my doctoral career without his support and help, and it is luckiest for me to pursue a PhD degree under his supervision. I am extremely grateful for his patient guidance, encouragement and selfless support during my doctoral life. He taught me how to start scientific researches and provided helpful advice about my research career. Whenever I met difficulties about academic or personal life, I can get instant help from him.

I would like to thank Dr. Yunchao Wei. He helped me a lot about my research and shared many valuable ideas with me. He also helped me to revise and promote the writings of my papers.

I would like to thank my friend Yu Wu for his help and kindness from the undergraduate period to the doctoral period. I would also like to thank my group members and colleagues, Linchao Zhu, Xin Yu, Pingbo Pan, Ping Liu, Xiaohan Wang, Peike Li, Qianyu Feng, Yutian Lin, Liang Zheng, Xiaojun Chang, Fan Ma, Zhedong Zheng, Hehe Fan, Xuanyi Dong, Yanbin Liu, Hu Zhang, Guang Li, Zhun Zhong, Yawei Luo, Qingji Guan, Guangrui Li, Qi Rao, Ruijie Quan, Tianqi Tang, Yang He, Zongxin Yang, Chen Liang, Xiaolin Zhang, Yuhang Ding, Yunqiu Xu, Youjiang Xu, and many others. Discussions about the research topics with them help me a lot.

I would like to thank my parents, Shiyang Miao and Xiufen Wang for their selfless love. They gave me any support when I needed it. They gave me strength and courage when I faced difficulties and challenges in my life.

I would like to thank my beloved wife, Danwen Sun for her sweet love. She is my best friend, my soulmate, and my heart. She always encourages me with her mild sound whenever I feel upset. She taught me to keep an optimistic attitude to life. Thanks for her patience and tremendous help.

# Abstract

With the development of deep learning, the deep models based on neural networks play an important role in vision applications. This dissertation focuses on two limitations of previous deep models. First, early approaches for vision tasks usually focus on global representations, while ignoring the discriminative partial features. However, partial representations provide sufficient recognition information for vision tasks and need to be well developed. Second, deep learning models are eager for massive data with labels, which is hard to acquire. The lack of labeled data inherently introduces uncertainty in deep models. Thus, a robust model should not only provide accurate predictions but also estimate uncertainty precisely.

This dissertation presents part-aware and robust deep models for some important vision applications, *i.e.*, the occluded person re-identification (re-id), the interactive video object segmentation (VOS) and the few-shot image classification. Concretely, for the occluded person re-id task, partial features are learned by partitioning the global feature map extracted by neural networks. Pose keypoints are adopted to indicate the visible and occluded parts. The information of occluded parts is depressed. For the interactive VOS, the partial similarity between adjacent frames is important to propagate segmented masks from the previous frame to the current processing frame. Thus, the pixel distances in a local part and the global map are computed for generating masks. For the few-shot image classification, a metric-based Bayesian framework is proposed for generating robust representations and reasoning about uncertainty, including calibration, recognition of out-of-distribution images and robustness against attacks.

In sum, I investigate the significance of the discriminative and robust partial representations and the ability of estimating uncertainty for the deep learning models, and apply them to some common vision applications to illustrate the effectiveness of the deep models.

# Table of contents

<b>List of figures</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.1.1 Part-Aware Deep Models for Vision Applications . . . . .	2
1.1.2 Robustness and Uncertainty in Deep Learning Models . . . . .	4
1.2 Thesis Organization . . . . .	4
<b>2 Literature Review</b>	<b>6</b>
2.1 Part-Aware Deep Models for Person Re-Identification . . . . .	6
2.2 Part-Aware Deep Models for Video Object Segmentation . . . . .	7
2.3 Probabilistic Deep Models for Few-shot Image Classification . . . . .	9
<b>3 Part-Aware Feature Learning for Occluded Person Re-Identification</b>	<b>11</b>
3.1 Introduction . . . . .	11
3.2 Occluded-Duke Dataset . . . . .	14
3.2.1 Properties of Occluded-Duke . . . . .	14
3.2.2 Collection of Occluded-Duke . . . . .	15
3.3 Keypoint-Guided Part-Aware Feature Alignment Model . . . . .	16
3.3.1 Preliminaries . . . . .	17
3.3.2 Visible Keypoints Generation . . . . .	18
3.3.3 Keypoint-Filtered Feature Branch . . . . .	19
3.3.4 Keypoint-Embedded Feature Branch . . . . .	19
3.3.5 Objective Function . . . . .	20
3.3.6 Part-Aware Feature Matching in Shared Visible Region . . . . .	21
3.4 Experiments and Analysis . . . . .	22
3.4.1 Datasets . . . . .	22
3.4.2 Implementation Details and Hyperparameters . . . . .	23

3.4.3	Evaluation Performance . . . . .	24
3.4.4	Ablation Studies . . . . .	26
3.4.5	Visualization . . . . .	29
3.5	Conclusion . . . . .	30
<b>4</b>	<b>Part And Whole Matching for Interactive Video Object Segmentation</b>	<b>32</b>
4.1	Introduction . . . . .	32
4.2	Memory Aggregation Networks with Part and Whole Matching . . . . .	34
4.3	Experiments . . . . .	40
4.3.1	Training . . . . .	40
4.3.2	Inference . . . . .	41
4.3.3	Segmentation Results . . . . .	43
4.3.4	Ablation Studies . . . . .	44
4.4	Conclusion . . . . .	46
<b>5</b>	<b>A Generic Bayesian Framework for Few-shot Image Classification</b>	<b>47</b>
5.1	Introduction . . . . .	47
5.2	Preliminaries . . . . .	48
5.2.1	Few-Shot Image Classification . . . . .	49
5.2.2	Natural-Gradient Variational Inference . . . . .	49
5.3	Metric-Based Bayesian Framework . . . . .	50
5.3.1	Objective Function . . . . .	51
5.3.2	Model Architecture . . . . .	52
5.4	Experiments . . . . .	54
5.4.1	Datasets . . . . .	55
5.4.2	Experiments Setting and Implementation Details . . . . .	55
5.4.3	Uncertainty Estimation . . . . .	56
5.4.4	Comparison about Adversarial Attacks . . . . .	59
5.4.5	Comparison about Overfitting . . . . .	60
5.4.6	Image Classification Accuracy Results . . . . .	60
5.4.7	Ablation Study . . . . .	60
5.5	Conclusion . . . . .	64
<b>6</b>	<b>Conclusion</b>	<b>65</b>
	<b>References</b>	<b>67</b>

# List of figures

3.1	Method [88] with global representations tends to generate error results. Images with green boundary are correct result while with red boundary are error results. . . . .	12
3.2	Difference between the partial (above) and occluded re-id settings (below). The partial re-id consists of the query set with obstacles and the gallery set with non-occluded images. The query images need to be cropped to remove the occluded parts. The occluded re-id's query set has obstacles and the gallery set has both obstacles and obstacle-free images. No pre-processing like manually cropping is needed. . . . .	13
3.3	Examples of the variations for occlusions. . . . .	15
3.4	The pipeline of our approach. We utilize red points to denote the <i>non-occluded</i> keypoints while green ones as the <i>occluded</i> keypoints. The proposed approach consists of three components. The Keypoint-Filtered Feature Branch utilizes the attentive maps generated by visible keypoints to remove occluded regions. The Keypoint-Embedded Feature Branch uses the non-occluded keypoints to produce the keypoint-embedding, which is utilized to re-weight the channel activations of the global feature map. The Part-Aware Feature Branch horizontally partitions the deep feature map to obtain the part-aware features. . . . .	17
3.5	Distance comparison strategy of our approach. The distances across images from probe and gallery sets are computed using part-aware features in the common non-occluded region as well as the keypoint-guided feature. . . . .	22
3.6	(a) The impact of the coefficient $\alpha$ . When $\alpha$ is 0, our model utilizes the keypoint-guided feature only. When $\alpha$ is 1, we adopt the part-aware features for evaluation. (b) Ablations on the part number $p$ . . . . .	28

3.7	(a) Ablations on the hyperparameters $\sigma$ of Gaussian maps. (b) Ablations on the image resolutions. . . . .	29
3.8	The appearance of the keypoint-masks produced by non-occluded keypoints. . . . .	30
3.9	Visualization of the outputs on the baseline [88] and the proposed approach. . . . .	30
4.1	An example of the turn-based interactive VOS. The green and red scribbles denote the true negative and false positive regions, respectively. At the first turn, the user provides green scribbles at frame 58. The model generates the segments of objects on the annotated frame and transfers the segments to other frames. In the following turns, the user refines the results by repeatedly drawing scribbles. For instance, at the second turn, the user draws green and red scribbles at frame 28. . . . .	33
4.2	The pipeline of our method, consisting of <b>pixel vector encoder</b> , an <b>interaction branch</b> and an <b>transference branch</b> . The pixel embedding vectors are extracted by this pixel vector encoder for each frame in one video. The interaction branch utilizes a “shallow” segmentation head to generate the segmentation results of the user-labeled frame. While the transference branch employs memory modules to accumulate the discriminative information and a segmentation head to predict segments of other frames. For the whole and partial maps, deeper green pixels have predictive results with higher confidence. . . . .	35
4.3	Matching operation on the whole map and local part. With regard to a pixel $p$ in the present frame (the $i^{th}$ frame), we compute distances between $p$ and pixels assigned to the object label in the interactive frame (whole map) by scribbles or the previous frame (partial map) by predicted segmentation mask. The nearest neighbor of the pixel $p$ in the embedding space is utilized to generate the matching map. . . . .	36
4.4	The scribbles are enhanced by computing the distance in the pixel vector space. . . . .	37
4.5	(a) <b>Memory module for whole matching maps</b> . Whole matching maps from the transference branch is accumulated as well as updated in this memory module. (b) <b>Memory module of the partial matching maps and the forgetting mechanism</b> . The partial matching maps in the transference branch is recorded. partial maps from the past $R$ turns are used to predict the segmentation masks. . . . .	37



4.6	At the first turn the scribbles are drawn only on the fore while no scribbles are on the back. To make our inference consistent for the beginning turn and the following turns, we employ a coarse region of interest (ROI) by assigning the pixels out of the region as the background.	41
4.7	The segment masks on DAVIS are visualized. Users' scribbles are synthesized by the computer, suggested by [7]. Segmentation masks are selected after 8 turns. . . . .	42
4.8	The impact of the proposed memory modules. All experiments are conducted on DAVIS. . . . .	45
4.9	Ablation studies on $T$ in the partial map memory module. $T$ indicates that partial maps in past $T$ turns in the memory module are utilized. .	45
5.1	Comparison between our approach and previous state-of-the-arts on the calibration curves and error scores. Bars closer to the diagonal line or lower ECE/MCE means better calibrated. Bars under the diagonal line or over the diagonal line indicates overconfidence or underconfidence, respectively. . . . .	57
5.2	The empirical CDF of entropies for the out-of-distribution examples. Ccloser to right-and-bottom means better uncertainty estimation. . . .	59
5.3	Comparison about the overfitting phenomena on FC100. The black and red lines denote the accuracy of the train and test set, respectively. . .	61
5.4	Comparison on the model calibration between BBB [5] and NGVI [36].	61
5.5	Impact of the Sampling Number on the accuracy (1st row) and the ECE score (2nd row). . . . .	63
5.6	(a) Comparison on the model calibration. (b) Comparison on out-of-distribution images. . . . .	64